MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE ON AWS

Lucian Revnic

AGENDA

Introduction

Infrastructure

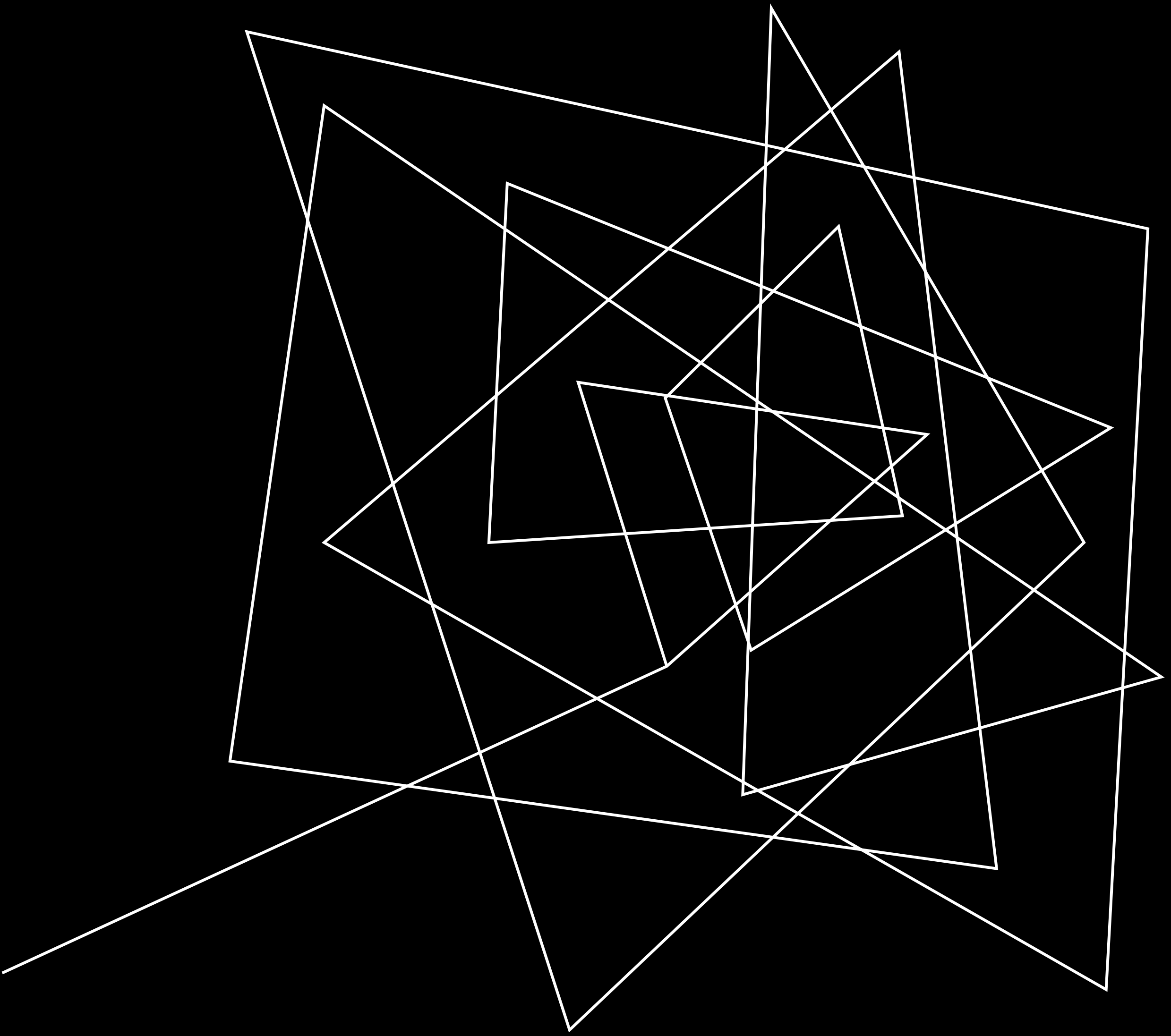
Tools

Applications and Services

INFRASTRUCTURE FOR TRAINING AND INFERENCE

BUILD AND DEVELOPMENT TOOLS

APPLICATION SERVICES

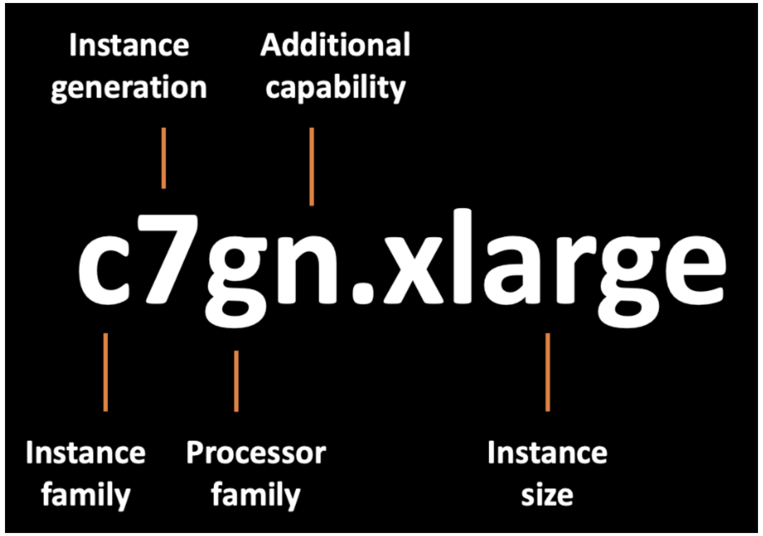


INFRASTRUCTURE

AWS INSTANCE TYPES – QUICK REFRESH

Instance families

- **C** – Compute optimized
- **D** – Dense storage
- **F** – FPGA
- **G** – Graphics intensive
- **Hpc** – High performance computing
- **I** – Storage optimized
- **Im** – Storage optimized with a one to four ratio of vCPU to memory
- **Is** – Storage optimized with a one to six ratio of vCPU to memory
- **Inf** – AWS Inferentia
- **M** – General purpose
- **Mac** – macOS
- **P** – GPU accelerated
- **R** – Memory optimized
- **T** – Burstable performance
- **Trn** – AWS Trainium
- **U** – High memory
- **VT** – Video transcoding
- **X** – Memory intensive



Instance generation

Additional capability

c7gn.xlarge

Instance family

Processor family

Instance size

Processor families

- **a** – AMD processors
- **g** – AWS Graviton processors
- **i** – Intel processors

Additional capabilities

- **d** – Instance store volumes
- **n** – Network and EBS optimized
- **e** – Extra storage or memory
- **z** – High performance
- **q** – Qualcomm inference accelerators
- **flex** – Flex instance

https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/instance-types.html

INFRASTRUCTURE FOR TRAINING AND INFERENCE

Amazon Elastic Compute Cloud (Amazon EC2)

EC2 P5

GPU base instances / GPU accelerated
8 NVIDIA® V100 Tensor Core GPUs
Up to 3200 Gbps net.
Deployed in EC2 **Ultra Clusters**
$$$$$

EC2 C5

Cheaper that GPU (P*, G*) instances
Intel or AMD processors
100 Gbps of net.
$

EC2 G5

Graphics intensive
NVIDIA A10G
Most cost-effective GPU instances inference & graphics
$$

EC2 Inf2

Reduce cost on **ML prod** deployments
Up to 12 AWS Inferentia2 accelerators
Up to 100 Gbps networking
$$

EC2 Tr1

Deep Learning Training
AWS Trainium chips
Intel Xel Scalable 3rd generation
Up to 16 AWS Trainium accelerators
Up to 1600 Gbps networking
Deployed in **EC2 Ultra Clusters**
$$$

AWS Deep Learning AMIs

x86 or AWS Graviton
Processor :GPU vs CPU vs Inferetia vs Hababa
SDK: CUDA vs AWS Neuron vs SynapsesAI
OS: Amazon Linux vs Ubuntu

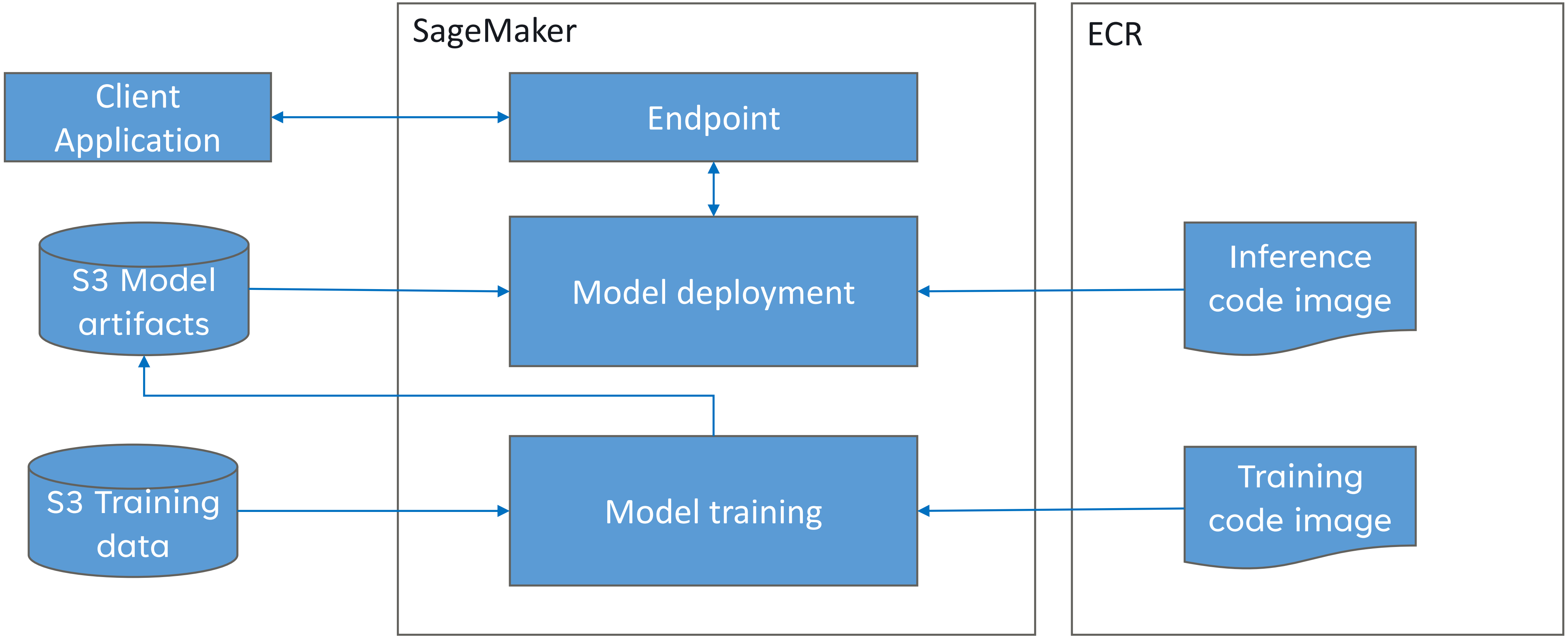
AWS Deep Learning Containers

60 DLCs for 5x frameworks (TensorFlow, PyTorch, MXNET, AutoGluon, Hugging Face)
EC2, ECS, EKS, Fargate, Lambda
General containers and AWS Neuron containers
For AWS or on-premises (eg ECS Anywhere, EKS Anywhere)

BUILD ML MODELS SERVICES AND TOOLS

AMAZON SAGEMAKER (1)

Fully managed machine learning (ML) service to build, train, and deploy ML models



https://www.sundog-education.com/

AMAZON SAGEMAKER (2) – BUILD & TRAIN

Built-in Algorithms

Supervise learning / Unsupervised learning
Text analysis/ Image processing /

Automatic Model Tuning

Identify best values for hyperparams

SageMaker Studio / Studio Classic

Create and share ML workflows

Jupyter Lab

Managed Jupyter Labs

SageMaker Studio Lab

Open Source / SageMaker Studio

SageMaker Debugger

Saves model state at periodic interval

SageMaker Autopilot /Auto ML

Automates algorithm selection (no code)

SageMaker JumpStart

Built-in FMs & alg.: AI21 Labs, HuggingFace, Alexa, Meta, co:here

SageMaker Clarify

Detect bias in your model

AMAZON SAGEMAKER (3) – DEPLOY BASED ON USE CASES

Source: https://www.softwebsolutions.com/resources/guide-to-amazon-sagemaker.html

Real Time Inference

Persistent, real-time endpoints

Serverless Inference

Workloads that have idle periods

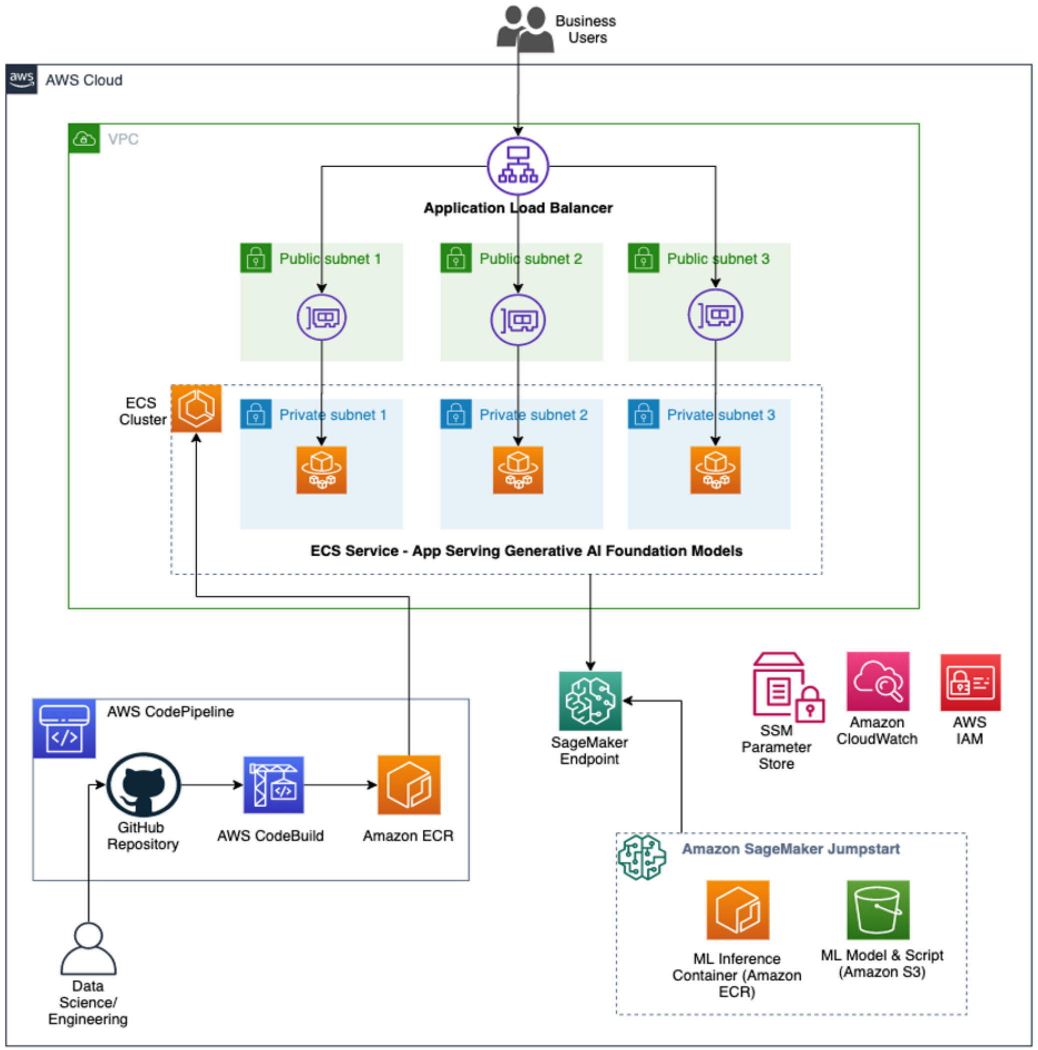
Asynchronous Inference

Large payloads (eg up to 1GB)

Batch Transform

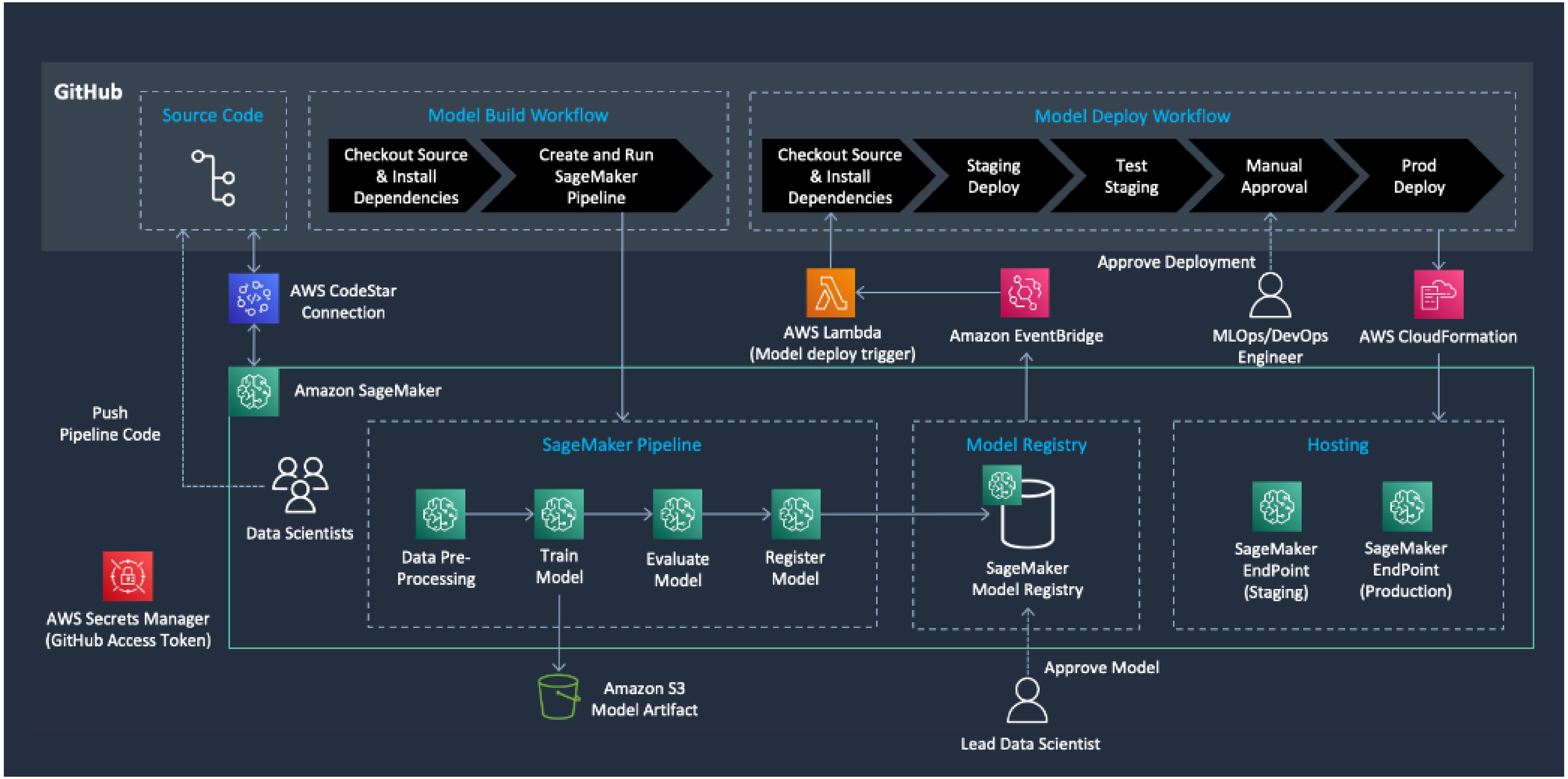
Prediction on entire data set

AMAZON SAGEMAKER – EXAMPLES (GENERATIVE AI APP)



https://aws.amazon.com/blogs/containers/build-generative-ai-apps-on-amazon-ecs-for-sagemaker-jumpstart/

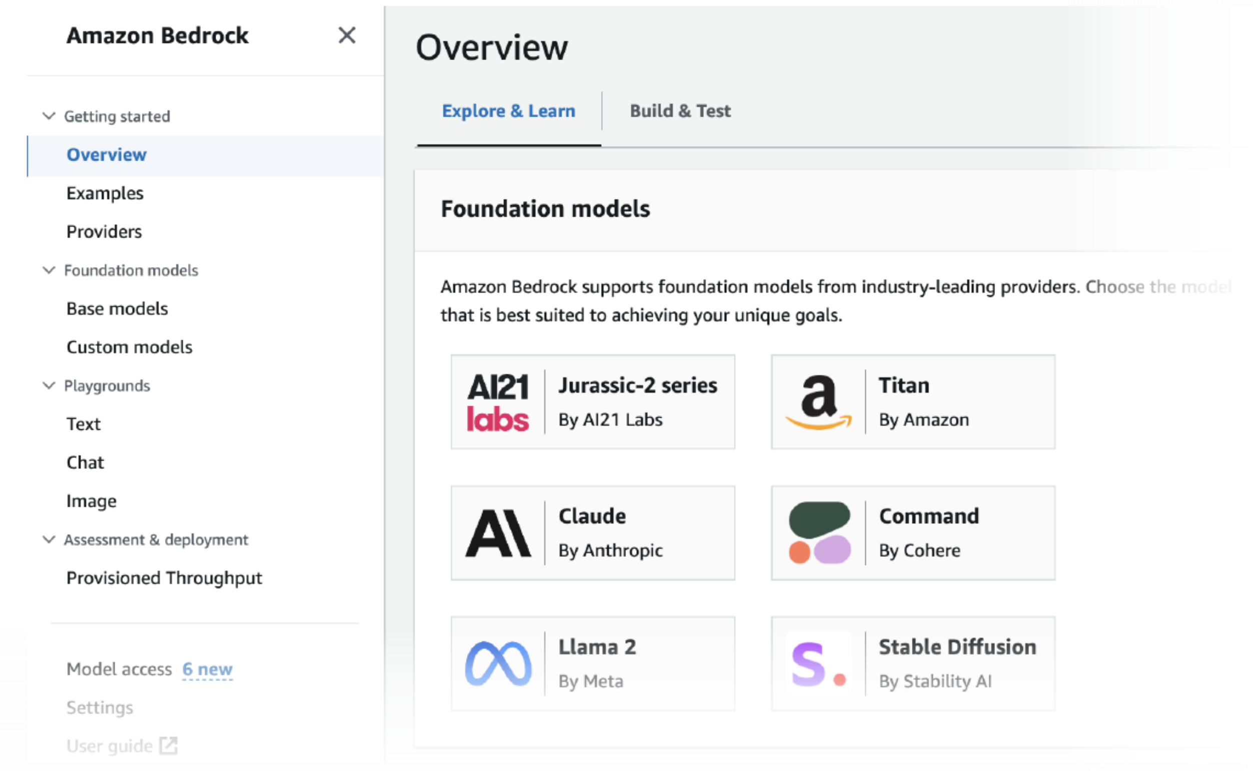
AMAZON SAGEMAKER – EXAMPLES (MLOPS PIPELINE)



https://aws.amazon.com/blogs/machine-learning/build-an-end-to-end-mlops-pipeline-using-amazon-sagemaker-pipelines-github-and-github-actions/

AMAZON BEDROCK

The easiest way to build and scale generative AI applications with foundation models



Amazon Bedrock ✕

Getting started

Overview

Examples

Providers

Foundation models

Base models

Custom models

Playgrounds

Text

Chat

Image

Assessment & deployment

Provisioned Throughput

Model access 6 new

Settings

User guide

Overview

Explore & Learn | Build & Test

Foundation models

Amazon Bedrock supports foundation models from industry-leading providers. Choose the model that is best suited to achieving your unique goals.

Jurassic-2 series By AI21 Labs	Titan By Amazon
Claude By Anthropic	Command By Cohere
Llama 2 By Meta	Stable Diffusion By Stability AI

Platform for FMs as it enables you to select your preferred model

Single API for inference

Playgrounds

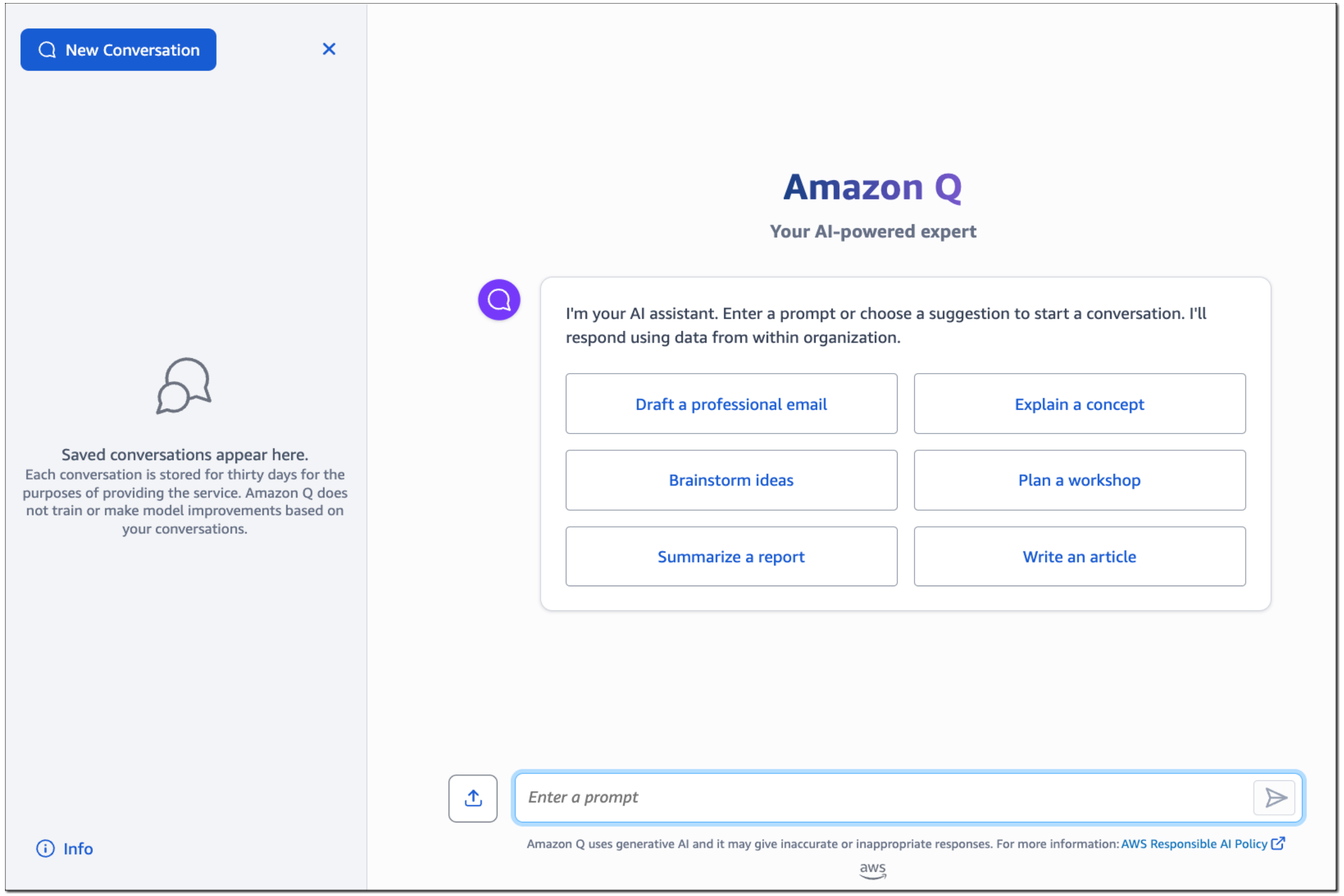
Source: https://aws.amazon.com/bedrock/developer-experience/

AI APPLICATION SERVICES

NEW!

AMAZON Q

The new generative AI–powered assistant



Saved conversations appear here.
Each conversation is stored for thirty days for the purposes of providing the service. Amazon Q does not train or make model improvements based on your conversations.

I'm your AI assistant. Enter a prompt or choose a suggestion to start a conversation. I'll respond using data from within organization.

Amazon Q uses generative AI and it may give inaccurate or inappropriate responses. For more information: AWS Responsible AI Policy

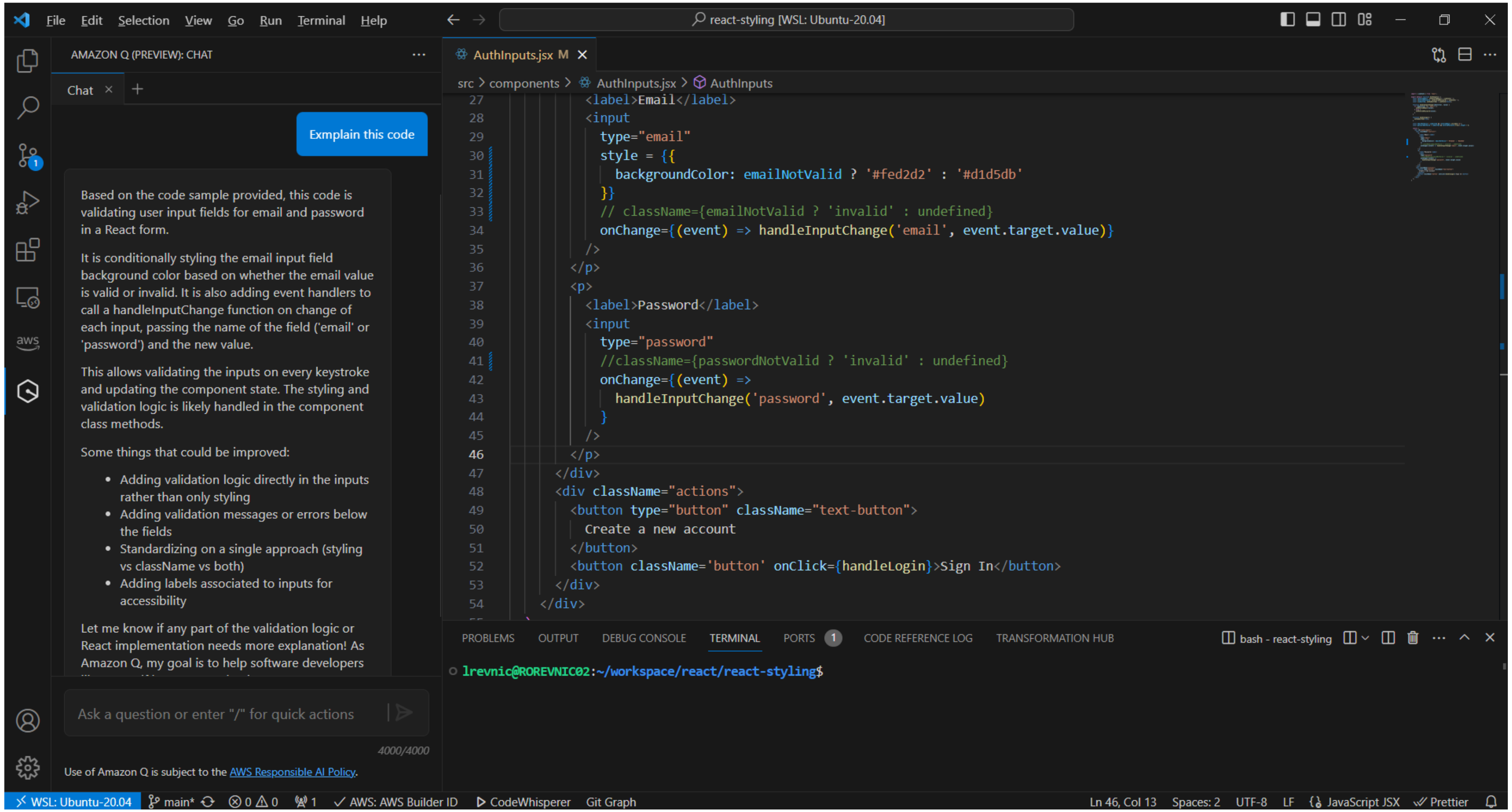
Connects to your data such as Jira, GitHub, ServiceNow

Integrated with AWS Console:
“Why is my instance not accessible?”

Integrated with AWS QuickSight:
“Build me a strategy to improve sales”

NEW!

AMAZON Q



Based on the code sample provided, this code is validating user input fields for email and password in a React form.

It is conditionally styling the email input field background color based on whether the email value is valid or invalid. It is also adding event handlers to call a handleInputChange function on change of each input, passing the name of the field ('email' or 'password') and the new value.

This allows validating the inputs on every keystroke and updating the component state. The styling and validation logic is likely handled in the component class methods.

Some things that could be improved:

- Adding validation logic directly in the inputs rather than only styling
- Adding validation messages or errors below the fields
- Standardizing on a single approach (styling vs className vs both)
- Adding labels associated to inputs for accessibility

Let me know if any part of the validation logic or React implementation needs more explanation! As Amazon Q, my goal is to help software developers

Use of Amazon Q is subject to the AWS Responsible AI Policy.

```
      <label>Email</label>
      <input
        type="email"
        style = {{
          backgroundColor: emailNotValid ? '#fed2d2' : '#d1d5db'
        }}
        // className={emailNotValid ? 'invalid' : undefined}
        onChange={(event) => handleInputChange('email', event.target.value)}
      />
    </p>
    <p>
      <label>Password</label>
      <input
        type="password"
        //className={passwordNotValid ? 'invalid' : undefined}
        onChange={(event) =>
          handleInputChange('password', event.target.value)
        }
      />
    </p>
  </div>
  <div className="actions">
    <button type="button" className="text-button">
      Create a new account
    </button>
    <button className='button' onClick={handleLogin}>Sign In</button>
  </div>
</div>
```

```
lrevnic@ROREVNIC02:~/workspace/react/react-styling$
```

Visual Studio Code Integration via AWS Toolkit

Amazon Q

Your generative AI-powered assistant tailored for work

AMAZON CODEWHISPERER

Code generation

Security scanning

IaC code generation (HashiCorp Terraform, AWS CloudFormation, CDK)

NEW!

OTHER ML AND AI SERVICES

Amazon Comprehend
NLP parsing and understanding

Amazon Elastic Inference
GPU Management - Deprecated

Amazon Forecast
Forecasting

Amazon Lex
Conversational interfaces

Amazon Transcribe
Speech to Text

Amazon Personalize
Recommendations

Amazon Polly
Text to Speech

Amazon Rekognition
Image recognition & video analysis

Amazon Textract
Document understanding / OCR

Amazon Translate
Machine translation

Amazon SageMaker Studio Lab

Amazon Comprehend Medical
Medical text and documents

AWS HealthOmics
(successor to Amazon Omics)

AWS Neuron
Deep Learning SDK

Amazon CodeGuru
Detect and FIX code vulnerabilities

Amazon DevOps Guru
Apps anomaly detection

AWS HealthLake
(successor to Amazon HealthLake)

AWS Panorama
Computer Vision

AWS HealthImaging
Medical Images in the Cloud

NEW!

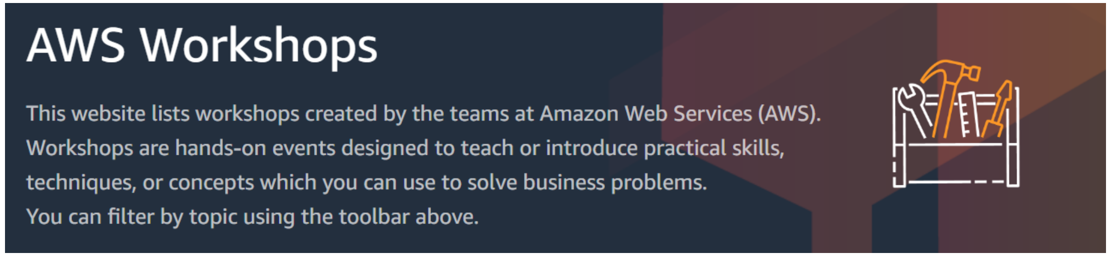
AWS HealthScribe

AWS HealthScribe
Clinical document generation

RESOURCES

AI/ML AWS Workshops

https://workshops.aws/categories/AI%2FML



AWS Workshops

This website lists workshops created by the teams at Amazon Web Services (AWS). Workshops are hands-on events designed to teach or introduce practical skills, techniques, or concepts which you can use to solve business problems. You can filter by topic using the toolbar above.

AWS Trainings – FREE!

https://skillbuilder.aws/

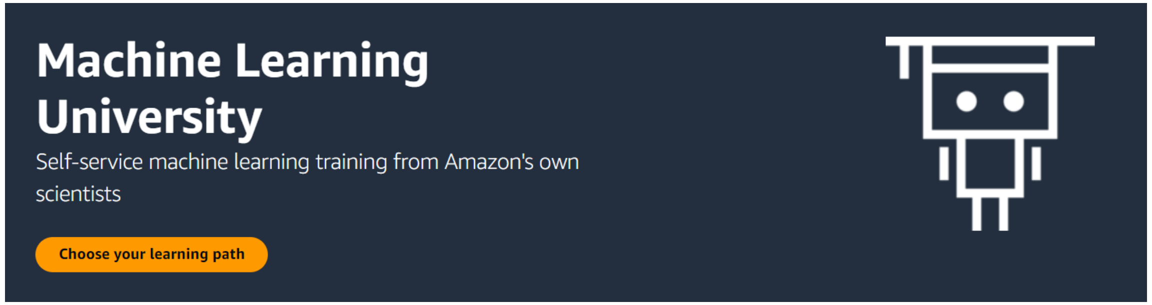
November 20, 2023

Amazon aims to provide free AI skills training to 2 million people by 2025 with its new 'AI Ready' commitment

Written by Swami Sivasubramanian, Vice President of Data and AI at AWS

AWS Machine Learning University

https://aws.amazon.com/machine-learning/mlu/



Machine Learning University

Self-service machine learning training from Amazon's own scientists

THANK YOU

Lucian Revnic

lrevnic@gmail.com

Transylvania Cloud

https://www.meetup.com/TransylvaniaCloud/